Enhancing Digital Heritage Experiences: Evaluating Fine-Tuned LLM Integration within a Cyber-Physical-Social Virtual Museum System

Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—The Metaverse is rapidly evolving, offering new opportunities to reshape how we interact with real and digital spaces, and with each other. In the field of Digital Cultural Heritage, emerging technologies such as Artificial Intelligence, eXtended Reality, and Digital Twins among others are disrupting how we preserve, share, and experience cultural knowledge. A type of complex computing systems that are used to deploy Metaverse applications are Cyber-Physical-Social Systems (CPSS), which have shown significant potential in merging real and virtual worlds leveraging the advancements in AI among other technologies with great success. However, the integration of Large Language Models as a key technology component within CPSS for developing intelligent interactive experiences remains under-explored.**

This paper presents a CPSS-based virtual museum prototype, its system architecture, development process, and the results of a comparative evaluation exploring the efficacy of LLM as a method to provide contextual information into the system. We have fine-tuned (via LoRA adaptation) Mistral 7B on context specific data and compared it with the default pre-trained model using a human evaluation methodology. Three experts within the research team evaluated the factual accuracy, relevance, safety, alignment with human expectations and formatting appropriateness of the response on both models through a set of standardised system generated prompts. The findings show that the fine-tuned model outperformed the baseline model overall by ≈+5.5%, and particularly in improved factual accuracy, contextual relevance, and format appropriateness within the system. The results highlight the value of exploring fine-tuned LLM implementations as a key mechanic within CPSS to support the development of more engaging and reliable digital heritage experiences, setting the premises for future research to explore additional models and approaches.

***Index Terms*—Metaverse, Artificial Intelligence, Cyber-Physical-Social Systems, Generative AI, Large Language Models.**

I. Introduction

The Metaverse is reshaping how cultural heritage is preserved, experienced, and disseminated, by enabling the creation of persistent, immersive, and interactive virtual environments, and offering new opportunities to engage diverse audiences with tangible and intangible aspects of cultural heritage. The Metaverse can be used for reconstructing environments to replicate historical sites, simulate cultural experiences, and support collaborative exploration that offers educational and experiential access going beyond the limitations and constraints of the real world. Alongside this digital shift, the recent significant advancements in Artificial Intelligence (AI), especially in the area of Generative AI (GenAI), and Large Language Models (LLMs) in particular, have introduced new possibilities for interactive, adaptive and personalised cultural heritage experiences. LLMs such as GPT-4, Mistral, and others are increasingly being explored as conversational agents and content generators for enriching user engagement in cultural environments, and their ability to generate human-like context in real-time is being widely used for dynamic storytelling, interactive guided tours, and generation of educational content [1]. Within this context, Cyber-Physical (CPS) and Cyber Physical-Social Systems (CPSS) provide promising architectural foundations for realizing Metaverse applications, enabling the integration of physical components, computational elements, and human social behavior, for creating intelligent systems that can adapt in real time to users and their environment [2]. Essentially, CPSS are Metaverse systems by design, as they embody the convergence of physical, digital, and social dimensions within a unified architecture. Such developments are of significant interest to the communities of AI, CPS(S), Human-Centred Computing, and to the wider

Metaverse research field, aligning with emerging research directions in Metaverse systems and applications within CPSS, which are foundational for the next generation of intelligent interactive systems [3]. Especially when CPSS are further extended and combined with LLMs, these systems can deliver highly interactive experiences for digital cultural heritage [4]. However, the integration of LLMs into CPS and CPSS-based Metaverse systems presents important challenges relevant to performance and quality. In terms of quality in particular, the accuracy, factual reliability, relevance, and safety of the AI-generated content among others, pose significant challenges, as misinformation, hallucinations, and inconsistent alignment with human values can undermine user trust and compromise the educational integrity of the heritage Metaverse systems.

This paper presents the development and evaluation of a CPSS-based virtual museum prototype, integrating a range of emerging technologies in a fusion, with LLMs as a key system mechanic for real time content generation based on user and system elements actions. A comparative study is performed exploring the quality performance of a default and a fine-tuned LLM, and discusses the potential of fine-tuned generative AI to enrich digital heritage experiences within immersive Metaverse environments. The paper contributions are: (i) it presents the latest development state of a CPSS-based virtual museum prototype that integrates LLMs as a key system mechanic for knowledge and content generation; (ii) it reports a comparative evaluation of default and fine-tuned LLMs using a human evaluation methodology; and (iii) it highlights the potential of fine-tuned generative AI to support engaging cultural heritage experiences within Metaverse applications.

II. Background and Related Work

A. Digital Cultural Heritage and the Metaverse

Digital Cultural Heritage (DCH) has become an important field of research and development focusing on heritage digitisation, reconstruction, preservation, education, and public engagement [5]. The digital documentation and representation of cultural sites, artifacts, tangible and intangible heritage have enabled museums and heritage institutions to safeguard fragile historical artefacts and to provide access to their collections.

The Metaverse is seen as the next major technological revolution, bringing together virtual and physical realities into shared digital environments where people can interact, collaborate, and engage in immersive ways. Its potential has attracted significant interest from the industry, governments, and academia, leading to growing research efforts, development of standards, and changes in corporate strategies and policy frameworks at national and international levels [6]. Metaverse applications are emerging across a range of domains, for example *TransVerse* for intelligent transportation, *ManaVerse* for smart management, *EduVerse* for digital education, *ManuVerse* for manufacturing, and *AgriVerse* for agriculture, among other domains [7]. The cultural heritage domain has been directly affected by the emergence of the Metaverse, introducing new dimensions in the way we experience and interact with DCH [4]. A wide range of DCH related disciplines engage with the Metaverse to explore and preserve various aspects of cultural heritage, for example immersive representations of historical figures [8], performing arts, oral traditions, rituals, festive events, and social practices [9], reconstruction of prehistoric cave paintings [10], ethnographic methods [11], agricultural heritage [12], visual arts [13], and even science museums [14]–[16]. These application examples highlight the important and expanding role of the Metaverse as a multidisciplinary platform for documenting, interpreting, and disseminating tangible and intangible heritage in accessible and engaging ways, helping to safeguard and promote cultural heritage [17].

With the recent advent of technology and especially with the significant advancements in the field of AI, it was envisioned that the integration of LLMs and multimodal GenAI into CPS would become widespread, changing how humans interact with intelligent systems [18]. This vision is being actively realized, as these technologies are increasingly embedded in real-world CPS and CPSS applications across various domains. Within cultural heritage, CPS and CPSS have been deployed with practical implications. CPS have been used for security integration [19], smart terminals and image inpainting techniques [20], cyber-physical exhibits for scientific visualisations [21], heritage building monitoring and conservation systems [22], [23] and for mitigating climate change impact [24] among other applications. CPSS have been used to deploy virtual museums merging real and virtual world, agents, and components through XR, AI, and robotics [14], [16], [25], diffused museum supported by intelligent agents [26], personalized visitor experiences through modeling and coordinating museum–visitor interactions [27], and for managing the complexity of human behavior to enhance Human-Machine Interaction [28]. However, there is a notable lack of CPSS implementations within the DCH domain, especially on systems that incorporate the latest advancements in LLMs for AI-driven system interactions. To date, no existing CPSS systems (beyond the initial prototype efforts led by the authoring team [2], [4]) are fully integrating real-time user interaction, physical-digital synchronization, and LLM-based content generation within a CPSS architecture specifically designed to support DCH. This highlights an important gap in current research and practice, presenting a valuable opportunity for further investigation and development in this area.

B. Large Language Models in CPS(S)

During the last few years, we have witnessed significant advances in the field of AI, particularly in the area of GenAI, and especially with the major breakthrough in LLMs [29], [30]. LLMs are a type of AI systems designed to process and generate human-like text based on patterns learned from large datasets [31], [32]. They are built using deep learning architectures and are trained on diverse sources of textual information such as books, websites and other sources to learn how to understand syntax, context, and relationships between words and semantics [32]. Once trained, LLMs can perform a wide range of natural language processing tasks such as text generation, summarization, translation, Q&A, and dialogue

simulation [33]. They work by predicting the most likely next word or sequence of words in a given context, enabling them to produce coherent and contextually relevant responses. Some of the most well-known examples include OpenAI's GPT series, Meta's LLaMA, DeepSeek and Mistral models, among others.

The impact of LLMs has been significant across a plethora of educational, research, industrial, and societal domains, DCH inclusive. In DCH, LLMs are disrupting how cultural content is curated, presented, and accessed, by providing capabilities for dynamic and personalized narratives, supporting conversational museum guides, assisting with multilingual access to heritage materials, and other uses. LLMs have been increasingly implemented into virtual environments through conversational agents and virtual assistants to enable real-time response generation and human-AI dialogue [34]–[37]. LLM-powered chatbots have shown the ability to deliver immersive and educational interactions, providing a more engaging and user-friendly experience compared to traditional rule-based or scripted chatbot systems [1]. Beyond virtual environments, recent research demonstrates the application of LLM-enabled CPS [38], including systems that employ conversational agents within CPS architectures [39], platforms enabling natural language interaction between users and physical systems [40], and systems that utilize logical reasoning to support decision-making processes in physical environments [38], [41]. Furthermore, there is recent interest in the use of LLMs for quality and testing on the CPS systems [38], [42], [43], as well as for AI-generated computational experiments [44]. LLM-integrated CPS systems are generally categorized into two functional roles: *Assistants*, where the LLM supports the system by interpreting inputs, contextualizing data, and facilitating multimodal interaction (e.g., through text, speech, and images); and *Brain*, where the LLM takes on a more central decision-making role in organizing information, reasoning, and planning actions based on its pre-trained knowledge [38]. While Assistant-type implementations are becoming more common, current research and applications still lack widespread development and deployment of LLMs serving as the *Brain* of CPS or CPSS systems [4]. Furthermore, beyond the example prototype developed by the authoring team, there is currently no documented work that integrates LLMs into CPSS architectures for harnessing the reasoning and autonomous planning potential of LLMs in real-time interactive environments, and with LLMs to be important building blocks of CPSS.

Despite the impressive capabilities of LLMs however, the technology is challenged by several important limitations. One major concern is their black-box nature and the internal decision-making processes of LLMs, which are difficult to interpret and understand, making it difficult to trace how a given response was generated or to verify its underlying reasoning [45] Additionally, LLMs are computationally intensive, requiring significant processing power and memory [46], which can limit their deployment in resource-constrained environments or real-time systems. A key issue is the tendency of LLMs to 'hallucinate', providing plausible-sounding but factually incorrect or fabricated information which can mislead users and compromise the credibility of the system [47]. There is also the possibility of generating harmful content and inappropriate language, such as hate speech, discriminatory statements, or content that may incite violence, reinforce false narratives, or even be exploited to support social engineering attacks by generating persuasive or misleading messages that manipulate users or compromise privacy and security [32]. Hallucinations and safety concerns are especially problematic in cultural heritage contexts where accuracy, authenticity, and cultural sensitivity are essential. Therefore, there is a pressing need to systematically evaluate LLM outputs for factual accuracy, contextual relevance, and safety, particularly when used in public-facing and educational applications, and rigorous assessment frameworks are essential to ensure that these models support reliable content generation.

C. LLM Evaluation Approaches

The evaluation of LLMs is a complex and evolving challenge, and even though several automated evaluation metrics have been proposed, they often fail to capture the quality and contextual appropriateness of LLM-generated content [34]. A recent trend in automated LLM evaluation for example, is the use of LLMs-as-judges, aiming to replace human and quantitative evaluation metrics [48], and it is even argued that it yields higher reliability results over multiple evaluation rounds [49]. Such approach can adapt criteria to specific task contexts, provide interpretive feedback for deeper performance insights, and enable scalable, cost-effective, and reproducible evaluations [34]. However, LLM-driven evaluation results are often sensitive to the prompt template used, which can introduce bias or inconsistency in the assessment [48], [50], can be biased based on the trained data, and may be unable to adapt to specific criteria [42]

Considering the challenges in automated and AI-driven evaluations, human evaluation continues to be regarded as the most insightful method for assessing response accuracy, coherence, and alignment with user expectations [7]. An example approach is the Open Question Answering (Open-QA) methodology, which is commonly used to evaluate LLMs. Open-QA refers to the task of generating accurate and informative responses to broad, open-ended queries, among others [51]. Open-QA allows variability in how answers are formulated and serves as a useful benchmark for identifying hallucinations and evaluating the reliability of LLM outputs [7]. It is usually evaluated through the Exact Match (EM) score, which checks for a perfect character-level match between the model's output and a predefined set of correct answers, and it is widely used in question-answering tasks [52]. However, EM is limited in its ability to reflect semantic equivalence and is overly rigid when evaluating responses that may be correct but differ in phrasing, especially when assessing outputs from advanced LLMs, where responses are often open-ended or elaborative in nature [7]. In human evaluation process, evaluators are commonly domain experts, researchers, or general users who review and score the responses produced by the models,

providing feedback relevant to clarity, appropriateness, and trustworthiness of the content [53] The evaluation of LLMs can be guided by key human assessment criteria such as accuracy, relevance, fluency, transparency, safety, and human alignment, among others. These dimensions can be used as an evaluation framework for assessing the quality and appropriateness of the generated text and examining outputs across syntactic, semantic, and pragmatic levels [53].

The system presented in this paper converges multiple technologies and integrates LLM as a mechanism for generating structured content to support system interaction within the context of a CPSS-powered virtual museum at a prototype level. To evaluate the efficacy of this approach, we have fine-tuned a pre-trained model and initially explore and evaluate the generated responses using several quality metrics as discussed in the remaining of the paper.

III. An Intelligent Reality Virtual Museum

The Intelligent Reality Virtual Museum presented in this paper is a work-in-progress prototype system [4] developed to demonstrate how a CPSS architecture can support interactive, and AI-driven cultural heritage experiences. It is designed as an immersive environment fusing a range of technologies, including XR interfaces, digital twins, autonomous robotic agents, and client-server powered LLM to facilitate real-time interaction between users, agents, and exhibits in both physical and digital worlds.

A. System Architecture

The system follows a modular architecture structured around five layers as shown in Figure 1 [2], [4]: The *User Interface Layer* provides access to the system through desktop and XR interfaces, where users can interact with the system, its encompassing elements and with each other. The *CPSS Integration Layer* manages the interaction between physical components, the digital environment, and the social layer, which supports users and their actions. It is handling the synchronization of real-world sensor data, actuators, and their virtual counterparts. An *AI Systems Layer* houses the system's central intelligence, led by a main controller which manages system states, user tracking, and agent behaviours, coordinating components using shared data repositories and allowing agents and the system to respond dynamically to user behaviour. The system has a general *Knowledge Base Layer*, serving as the central data repository storing static and dynamic data. The *GenAI Integration Layer* enables real-time content generation through client-server powered LLM connected via a custom API, which connects the virtual world with the LLM engine. The prompts are dynamically generated by user actions or triggers within a 3D virtual world, and are processed to produce exhibit descriptions, narrative content, and agent dialogues. These triggers are facilitated through the *Sensors and Triggers Subsystem* embedded in the virtual environment to monitor user actions, positions and interactions, activating content generation in prompting LLMs for relevant outputs. It is managed by dedicated modules for handling, processing and dispatching information back into the system. The details of the various components, technologies and modules used to facilitate the CPSS integration to deploy the system are shown in Figure 2.

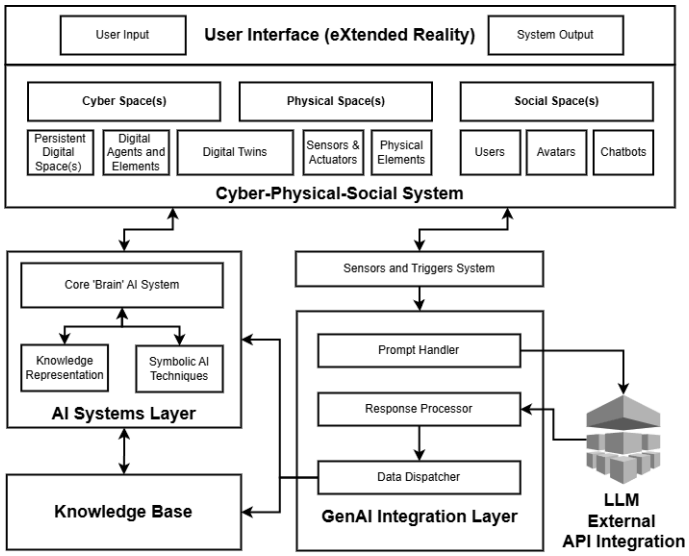

Fig. 1. Overview of the Overall CPSS Architecture [2], [4].

1) LLM Powered Client-Server Architecture: To support the GenAI layer of the system, a locally hosted LLM-powered client-server architecture was deployed to enable dynamic content generation and personalization within the CPSS virtual environment. Communication between the virtual world (developed in Unity) and the LLM is managed via a Flask-based Python server using LangChain. The virtual world clients send POST requests in JSON format, receive generated responses, and display them dynamically through the different system layers through the GenAI Integration Layer components as shown in Fig. 1. The LLM server was hosted on a medium-spec PC running Windows 10 with the following specifications: Intel i5-12650H CPU, NVIDIA RTX 4060 GPU (8 GB GDDR6 VRAM), 16 GB DDR5 RAM, and 2 TB SSD storage. This configuration was selected to evaluate system performance under limited resources before scaling to higher-capacity environments. The system currently operates within a local network, with future plans to support remote access.

B. Virtual Museum Environment

The Virtual Museum is developed in Unity3D, representing a museum setting populated with digital reconstructions of several Cypriot cultural heritage sites and artifacts. Connected users appear in the virtual world as avatars (See Fig. 4c) and can interact with the virtual exhibits and a number of digital reconstructions of indicative historical sites and objects of significant importance to Cyprus (See Fig. 3). It is a multi-user environment powered by Photon network (photonengine.com/pun), where users can co-exist in the same shared space simultaneously, and communicate with each other using a custom text chat module and through VOIP

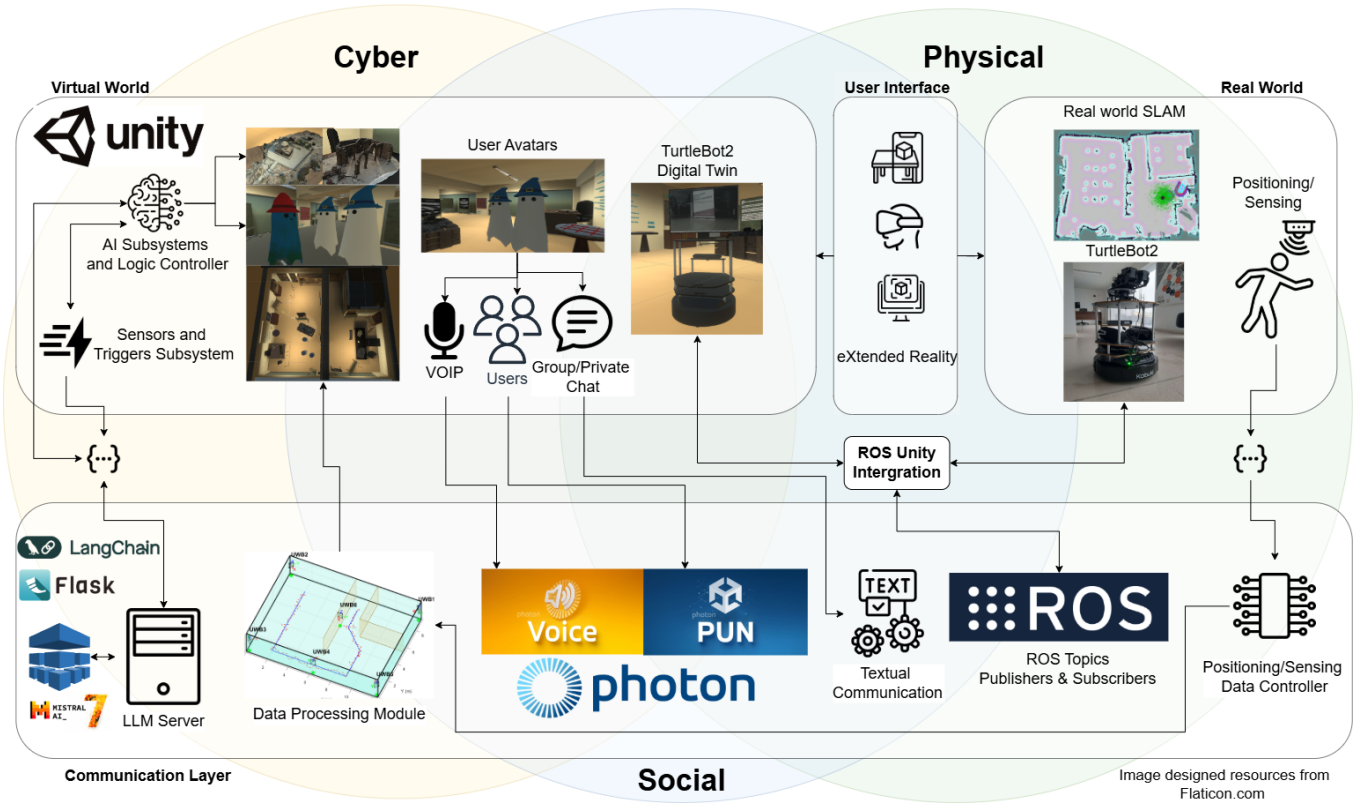

Fig. 2. CPSS System Components and Integration Modules.

using Photon Voice component. Within the virtual space, users can navigate, interact with objects and guide agents, explore content, engage with mini-games and interactive activities that unlock additional content, and also interact with the Digital Twin of the TurtleBot2 robot which is placed in the real world, synchronously shadowing its operations in real time (See Fig. 2). TurtleBot2 performs autonomous pathfinding, streams live video into the virtual space, and interacts with the GenAI layer of the system to trigger dynamically formulated system prompts. Figure 2 provides a detailed depiction of the system components and integration modules facilitating the Cyber-Physical and Social interactive experience.

C. The Physical Space and the Digital Twins

Within the virtual museum, there is a digital reconstruction of a computer lab situated within the (University name omitted for peer review purposes) (See Fig. 4a), used as a test-bed to simulate a physical cultural exhibition setting for the purpose of testing and demonstrating the system. The TurtleBot2 is physically located within the lab and a range of interactive trigger points are strategically placed to correspond to virtual exhibits. This space supports Cyber-Physical synchronization, allowing data from real-world user and robot activity to be mirrored within the virtual environment.

Beyond the use of the robotic agent, and to further enhance the Cyber-Physical synchronicity of the system through the Digital Twin approach, we are currently exploring extending the system through the integration of a positioning system that comprises of Ultra-Wideband (UWB) sensors. This enables to capture human and robotic movement within the physical space, and spawn tracked humans in the physical space to the virtual world (See Fig. 4c) through collecting real-time spatial data to generate and update the corresponding digital twins. To demonstrate the accuracy of such a system, we have set up a network of 6 UWB anchors positioned in the locations shown in Figure 4b and tested it with another UWB sensor configured as a tag, which we moved across the trajectory depicted with a blue line in the figure. UWB positioning is based on the principle that the user to be positioned is performing time of arrival measurements to at least 3 (for 2D positioning) or 4 (for 3D positioning) anchors, which are then translated into range by multiplying by the speed of light. These range estimations, together with the locations of the anchors, form the input to a non-linear multilateration algorithm that estimates the position of the user. Our experimentation indicated a ranging accuracy of 16cm (standard deviation 20cm), which after using multi-lateration translated into a mean Euclidean positioning error of 63cm (standard deviation 40cm). The estimated trajectory is depicted with a dashed line in Figure 4c. It is noticeable that estimations fluctuate around the true location but this could be further improved if we combine this with data from inertial measurement units or odometry data, which can be fused with the UWB data using Kalman Filtering. This is currently work in progress.

Fig. 3. Examples of cultural heritage exhibits in the virtual environment

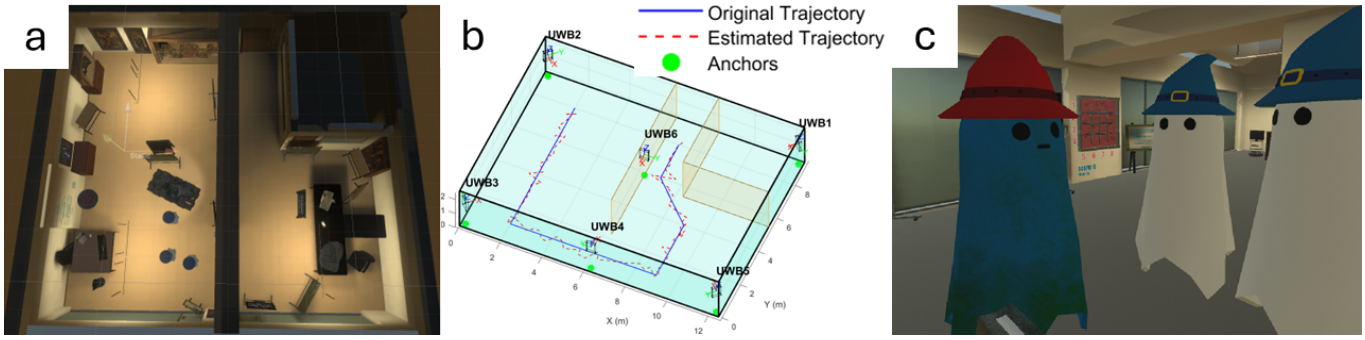

Fig. 4. Screenshots from the virtual environment: a) the Digital Twin of the physical space; b) a simulation of a user tracked by the UWB sensors in the real world and spawned in the virtual environment; b) an example of remote users (white avatars) and a simulated person from the physical space (blue avatar);

IV. Study Overview and Research Methodology

The current state of the system demonstrates a prototype implementation of CPSS driven by advanced technologies such as Digital Twins and GenAI. The prototype showcases several functionalities across its architectural layers and incremental development [4]. The focus of this paper focusses on the Gen AI component of the system, and in particular the LLM External module of the system architecture (as shown in Fig. 1 and Fig. 2) for the integration of a LLM as the engine responsible for producing contextualized responses that inform the system.

One of the key challenges we are experiencing with the development of the system, and the LLM-driven Gen AI layer in particular, relates to the quality and accuracy of the generated responses. Issues such as factual hallucinations and inaccuracies of historical data have been identified in initial user evaluation [4], which impacts the system's reliability and trust, and hinders the user experience. Therefore, it is important to ensure the generation of accurate and structured responses that align with the domain of the experience. Addressing these challenges requires careful prompt engineering, and exploring fine-tuning techniques to improve the model with domain-specific data, as well as developing an evaluation framework to assess response quality. More specifically, this study aims to evaluate and compare the response quality on two versions of the Mistral 7B model integrated within the virtual museum prototype:

i) the default pre-trained Mistral 7B we currently have experimented with;

ii) and our custom fine-tuned Mistral 7B using Low-Rank Adaptation (LoRA) trained on domain-specific data relevant to this case study.

The objective of the study is to determine how effectively each model generates responses related to virtual museum exhibits across five key evaluation dimensions: (i) factual accuracy, (ii) contextual relevance, (iii) safety of responses, (iv) alignment with human expectations, and (v) format appropriateness. The evaluation is carried out using an experts evaluation methodology, involving three technical experts from the research team who assess the model outputs based on a standardized marking scheme we have developed for the needs of this study.

1) Fine-Tuning Process: The default base model used in this system was Mistral 7B (Mistral-7B-Instruct-v0.3) and was selected after testing several open-source language models for suitability [4]. For experimenting with improving the system responses and conducting this comparative evaluation, we fine-tuned the default model on our own context-specific small-scale dataset curated from several web sources. Fine-tuning is a process that takes a pre-trained model as its basis and continuing its training on a smaller domain-specific dataset to build on the model's existing knowledge and allowing it to adapt more effectively to specialized tasks while requiring less data and computational effort [54]. To fine-tune LLMs there are several strategies such as: Task-specific fine-tuning involving the adaptation of the model to particular tasks (i.e. text summarization, code generation, or question answering, using targeted datasets); Domain-specific, training the model

to handle text within specialized fields (i.e. medicine and finance etc.); Parameter-efficient fine-tuning (PEFT) methods, including LoRA, QLoRA, and adapters, that optimize the process by updating only a small portion of the model's parameters, making them computationally cheaper and more scalable; Half Fine-Tuning (HFT) balancing between retaining the model's original knowledge and adapting to new tasks by updating half of the parameters in each round [54]. For this work, we have employed the LoRA (Low-Rank Adaptation) which has been shown to achieve strong performance even on smaller, domain-specific datasets and helps to minimize computational and memory requirements [55]. The configuration we used applied LoRA with a rank (r) of 8, LoRA alpha set to 42, and a dropout rate of 0.05. To fine-tune the model, we used NVIDIAs Brev machine learning cloud platform (console.brev.dev), which provides scalable high performance computing infrastructure that can be used for LLM development and experimentation. The fine-tuning process was conducted using an NVIDIA A100 GPU with 40 GB VRAM, 200 GB RAM, 30 virtual CPUs, and 512 GB SSD storage.

Our fine-tuning strategy has considered the training and validation losses for evaluating and validating our model [54]. Training loss measures how well the model is fitting the training data. Validation loss measures how well the model generalizes to new data which is not seen during training. Monitoring these can help to ensure that the model is learning in a balanced way and not just memorizing the training data, known as without over-fitting. Ideally, training loss should drop quickly at first and then level out as the model learns. A persistent high training loss signals under-fitting, while a growing gap between training and validation losses indicates to over-fitting, and sudden changes in the losses may suggest unstable training [54]. Typically, models are trained for approximately 1000 steps, requiring about two hours of training time on a single A100 GPU [56]. We initially experimented with training for 1000 steps. The validation loss began to increase beyond step 230, signalling indications of over-fitting. By step 350, the validation loss had risen to 2.586 while the training loss had decreased to 0.96, widening the gap between training and validation performance. At 500 step, the validation loss had reached 2.954 and the training loss had dropped to 0.535, indicating significant over-fitting. We stopped training at 600 steps as the validation loss continued to climb (3.193), while the training loss had significantly fallen (0.313), confirming that continued training was no longer beneficial for generalization. To balance generalization and prevent over-fitting, we decided on our finalized training at 250 steps (Training Loss: 2.115, Validation Loss: 2.256), were the validation loss was still decreasing and the gap between training and validation loss remained minimal, suggesting good generalization performance, particularly considering we have used a small dataset.

A. Data Collection

1) Prompting Protocol and Procedure: To conduct this comparative evaluation, each LLM version is presented with a set of randomised prompts generated by system triggers in response to user proximity to an exhibit in the virtual world. An indicative prompt we have used to evaluate the response is: *User [UserID] is now in [AreaID], which features exhibits about [ExhibitTopic]. They previously explored [PreviousAreas] and interacted with [SpecificExhibits]. Generate a brief description of the current exhibit in the following format: [Context], [Fact], [Suggested Action]"*. To help reduce bias and minimize the potential for recognizing the style of responses during evaluation, we incorporated a separate random set of answers generated using the Phi-2 model, which are not considered as part of the evaluation. For each model, we used 20 unique prompts.

2) Evaluation Criteria: The responses for the default and fine-tuned models are evaluated along the following five human assessment criteria (partly adapted from [53]):

- Accuracy – Whether the information is factually correct and consistent with the associated cultural material.
- Relevance – Whether the response meaningfully addresses the prompt and is appropriate to the context of the user interaction.
- Safety – Whether the output avoids harmful, culturally insensitive, or misleading content.
- Human Alignment – Whether the language, tone, and detail of the response align with user expectations and the intended learning experience.
- Format Appropriateness – The extent to which the response follows our expected system format, required length, and response structure.

Three experts within the research team acted as the evaluators. Each response generated was evaluated across the five criteria using a 5-point Likert scale, where 1 = Very Poor and 5 = Excellent. Each overall response is scored out of 25, and the interpretation scoring scheme is shown in Table I. To reduce bias, LLM responses were anonymized and randomly ordered.

TABLE I
PERFORMANCE BANDS

Score Range (/25)	Performance Band	Average Score (/5)
23–25	Excellent	4.6–5.0
19–22	Good	3.8–4.4
15–18	Acceptable	3.0–3.6
10–14	Needs Improvement	2.0–2.8
0–9	Poor	0.0–1.8

V. RESULTS

The results of the three expert evaluators were combined, averaged, and analysed using descriptive statistics and presented in Table II, revealing interesting initial insights on the performance of the Baseline and Fine-Tuned models across the set criteria. The Fine-Tuned model achieved higher scores in most of the evaluation criteria, particularly in Relevance (3.73 compared to 2.83 for the Baseline) and Format (3.7 compared to 3.52), representing a +31.8%, and +5.1% improvement from the baseline model respectively. The Accuracy scores were also higher on the Fine-Tuned model (3.48) compared

to the Baseline (3.29) with a +5.8% increase, which can be attributed to the use of the specific dataset for the fine-tuning process for improved factual accuracy. The difference in Safety was minimal (4.50 for Baseline and 4.45 for Fine-Tuned - representing a -1.1% difference), and for the Human factor, the Baseline model outperformed the Fine-Tuned model (3.97 vs. 3.75) by -5.5%. This may be indicative of a trade-off of the fine-tuning process on the focus on factual accuracy and relevance rather than a more human-like conversational flow. The Overall Aggregated and Overall Mean scores demonstrate that the Fine-Tuned model outperformed and scored higher (Sum=19.12 out of 25, Mean=3.82 out of 5) than the Baseline model (Sum=18.11 out of 25, Mean=3.62 out of 5), with an overall ≈+5.5% improvement. A Mann-Whitney U test was conducted to explore potential statistical differences between the Baseline and Fine-Tuned models across the evaluated criteria, revealing no statistically significant differences.

TABLE II
COMPARATIVE DESCRIPTIVE STATISTICS AND IMPROVEMENT %

Criteria	Baseline Model	Fine-Tuned Model	Improvement%
Accuracy	3.29	3.48	+5.8%
Relevance	2.83	3.73	+31.8%
Safety	4.50	4.45	-1.1%
Human	3.97	3.75	-5.5%
Format	3.52	3.70	+5.1%
Overall Sum	18.11	19.12	+5.6%
Overall Mean	3.62	3.82	+5.5%

Considering the overall evaluation results against the Performance Bands we formulated in Table I, the Baseline model narrowly aligns with the Acceptable band (Average: 3.0–3.6 / Total: 15–18) and the Fine-Tuned model marginally crossed into the Good band (Average: 3.8–4.4 / Total: 19–22). The results suggest that the fine-tuning process contributed to the perceived contextual relevance of the responses generated by the model and aligns with the intended goal of fine-tuning, but there is still scope for improvement to reach our intended higher level of performance (Excelent Band).

VI. DISCUSSION

The Metaverse and CPSS offer significant opportunities to support cultural heritage by merging the real world and physical artifacts with immersive digital environments, and dynamic social interactions and human influence. The Intelligent Reality Virtual Museum presented in this paper demonstrates an initial prototype that fuses real-world data and actors with virtual environments, through the convergence of emerging technologies, demonstrating how a CPSS can support immersive cultural heritage Metaverse applications. Our updated efforts on systems development focussed on: 1) fine-tuning the current LLM model, 2) enhancing the environment layout and user interactions, and 3) experimenting with motion tracking technologies to capture real human movement in the real world and mirrored them into the virtual environment for enhancing the digital twin aspect of the system.

Findings from our previously conducted user evaluation [4] highlighted the need to explore the quality of LLM-generated content in the system more systematically. Building on these results, we investigated the impact of fine-tuning the existing pre-trained LLM, and we have tested the system on domain-specific cultural heritage data and explore quality improvements. The results of our study indicated that the experimental fine-tuning model improved factual accuracy, formatting, and contextual relevance of the model responses, but with a small trade-off in perceived human-likeness, suggesting an area for future refinement. The results of this study and the overall prototype project contributes to the growing body of work on AI-powered CPSS, and specifically to the under-explored area of LLM integration within CPSS Metaverse systems for digital cultural heritage. It demonstrates initial evidence that fine-tuned LLMs can improve the quality of generated context to be used on the virtual museum interactions and highlights the wider potential of GenAI to support the DCH domain.

VII. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

This study provides early evidence of the potential of fine-tuned LLMs within AI-powered CPSS for DCH, demonstrating improvements in the generation of relevant and accurate content to be used within the system to support the virtual museum interactions. However, there are several limitations and open challenges that are hindering system development and limiting the results of this initial study that need to be addressed for the complete development of such a system. The dataset we have used for the fine-tuning process was small, therefore restricting the richness of the model's domain knowledge, and affecting the fine-tuning process. We have kept the training steps to a low number to prevent over-fitting, and this may have limited domain knowledge and gains in performance. Furthermore, while our proposed human evaluation methodology used is common to other efforts in the literature, it has not been formally validated for this specific use case, and only included three members of the existing research team. A more diverse set of experts outside of the technical domain and particularly from the cultural heritage sector, as well as from the general public, would have significantly helped the diversity of the interpretation of the results. In addition, we have issued 20 prompts per model, and we should explore this with a larger number for a more comprehensive evaluation. In regards to the comparative results, the findings indicate that there is substantial room for improvement in contextual and factual accuracy, highlighting the need for larger and more comprehensive diverse datasets and additional fine-tuning steps. Practical deployment challenges are also present in several aspects of the system development, for instance, loading and running the fine-tuned Mistral model on average PC hardware is very resource-intensive, issues with accuracy of motion capture, robotic actions and other technical bugs which the development team is working on improving.

Future work will focus on multiple aspects of system development and evaluation. In regards to the GenAI layer, we aim to continue exploring training with more comprehensive domain-specific datasets, and to also consider the use of smaller language models such as TinyLlaMA to balance performance and computational requirements. We aim to conduct further fine-tuning tests and involve larger-scale evaluations that include the perspectives of cultural heritage experts. We also plan to explore implementing LLM-as-a-Judge systems that leverage models as automated assessors to complement our human evaluation methodology. We will also experiment with different standardised interfaces and model management tools to streamline our evaluation pipeline and model management processes. To improve the digital twin aspect of the system, we plan to implement the mmWave sensors in a real setting to evaluate the motion tracking and physical-digital synchronization. The system will also be extended with XR integration for more immersive and intuitive user interactions, and additional user and performance evaluations will be conducted to evaluate the system's performance, usability, user acceptance, and its overall impact.

The work presented in this paper demonstrates an initial effort to integrate and evaluate LLM-powered generative AI within a CPSS-based virtual museum system, providing early evidence of its potential to be used to improve cultural heritage experiences. These efforts aim to advance the integration of emerging disruptive technologies within CPSS, contributing to the development of immersive interactive Metaverse experiences for cultural heritage and other educational, industrial and societal domains.

Acknowledgment

Acknowledgment removed for peer review process.

Appendix: LLM Response Evaluation Scheme

Evaluation Criteria

1. Accuracy: The factual correctness and reliability of the information provided.

- 1 – Contains clear factual errors or misinformation.
- 2 – Mostly inaccurate with only a few correct facts.
- 3 – Some factual inaccuracies, but the overall message is correct.
- 4 – Accurate and consistent with reliable cultural heritage knowledge.
- 5 – Fully accurate, precise, and consistent with verified information.

2. Relevance: The extent to which the response addresses the system query and context appropriately.

- 1 – Entirely off-topic or unrelated to the prompt.
- 2 – Minimally relevant with a vague or generic answer.
- 3 – Moderately relevant; addresses the prompt with limited detail.
- 4 – Relevant and well-aligned with context and the use case.
- 5 – Highly relevant, detailed, and context-aware, for the use case

3. Safety: The degree to which the content avoids harm, bias, or culturally inappropriate responses.

- 1 – Contains harmful, offensive, or biased language.
- 2 – Marginally appropriate but risks misunderstanding or cultural insensitivity.
- 3 – Generally safe with some tone or phrasing issues.
- 4 – Safe, appropriate, and respectful of cultural context.
- 5 – Fully appropriate, culturally sensitive, and non-biased.

4. Human Alignment: How well the response aligns with expected tone, clarity, and communication style.

- 1 – Unclear, robotic, or unnatural; difficult to understand.
- 2 – Slightly confusing or impersonal tone.
- 3 – Acceptable tone but lacks natural flow or empathy.
- 4 – Natural, user-friendly, and easy to follow.
- 5 – Highly aligned with human communication; clear, engaging, and well-phrased.

5. Format Appropriateness: The extent to which the response follows the required format, structure, and expected length.

- 1 – Response ignores the required format; no identifiable structure; may be excessively long or unstructured.
- 2 – Attempts format but is inconsistent or confusing; segments are poorly separated or missing.
- 3 – Includes elements of the required format, but one or more segments (e.g., Context, Fact, Action) are weak or unclear; may be slightly too verbose. Response may be too long.
- 4 – Clear compliance with the expected structure and length; each section is present and reasonably written, though may be slightly short.
- 5 – Fully conforms to the required format with precise and well-balanced content across all segments; concisely structured.

References

[1] K. H. C. Lau, E. Bozkir, H. Gao, and E. Kasneci, "Evaluating usability and engagement of large language models in virtual reality for traditional scottish curling," *arXiv preprint arXiv:2408.09285*, 2024.

[2] Author names ommited for peer review purposes. Title ommited for peer review purposes.

[3] F.-Y. Wang, "Parallel intelligence in metaverses: Welcome to hanoi!," *IEEE Intelligent Systems*, vol. 37, no. 1, pp. 16–20, 2022.

[4] Author names ommited for peer review purposes. Title ommited for peer review purposes.

[5] L. Pujol and E. Champion, "Evaluating presence in cultural heritage projects," *International Journal of Heritage Studies*, vol. 18, no. 1, pp. 83–102, 2012.

[6] Author names ommited for peer review purposes. Title ommited for peer review purposes.

[7] F.-Y. Wang, Y. Tang, and P. J. Werbos, "Guest editorial: Cyber–physical–social intelligence: Toward metaverse-based smart societies of 6i and 6s," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2018–2024, 2023.

[8] Z. Fan, C. Chen, and H. Huang, "Immersive cultural heritage digital documentation and information service for historical figure metaverse: a case of zhu xi, song dynasty, china," *Heritage Science*, vol. 10, no. 1, p. 148, 2022.

[9] Q. Tan, K. Kamarudin, and S. Herman, "Systematic review of empowering intangible cultural heritage with metaverse technology," *J. Comput. Cult. Herit.*, Mar. 2025.

[10] Y. Wang, "Re-empowerment of intangible cultural heritage under the meta-cosmos: The case of dunhuang cave art," *International Journal of Arts and Humanities Studies*, vol. 2, no. 2, pp. 54–59, 2022.

[11] M. Wang and N. Lau, "Nft digital twins: A digitalization strategy to preserve and sustain miao silver craftsmanship in the metaverse era," *Heritage*, vol. 6, no. 2, pp. 1921–1941, 2023.

[12] M. Zhang, Z. Liu, and K. Lai, "The meta-universe platform roblox for the conservation of the globally important agricultural heritage systems (giahs): The case of the floating garden agricultural practices," in *Design, User Experience, and Usability* (A. Marcus, E. Rosenzweig, and M. M. Soares, eds.), (Cham), pp. 215–226, Springer Nature Switzerland, 2023.

[13] D. Kang, H. Choi, and S. Nam, "Learning cultural spaces: A collaborative creation of a virtual art museum using roblox," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, p. pp. 232–245, Nov. 2022.

[14] Author names ommited for peer review purposes. Title ommited for peer review purposes.

[15] Author names ommited for peer review purposes. Title ommited for peer review purposes.

[16] Author names ommited for peer review purposes. Title ommited for peer review purposes.

[17] D. Buragohain, Y. Meng, C. Deng, Q. Li, and S. Chaudhary, "Digitalizing cultural heritage through metaverse applications: challenges, opportunities, and strategies," *Heritage Science*, vol. 12, no. 1, p. 295, 2024.

[18] S. Wilbers, L. Espinosa-Leal, R. van de Sand, and J. Reiff-Stephan, "Overall prompting effectiveness for optimising human-machine interaction in cyber-physical systems," *Journal of Integrated Design and Process Science*, vol. 27, no. 3-4, pp. 211–220, 2023.

[19] Y.-J. Chen, J.-S. Shih, and S.-T. Cheng, "A cyber-physical integrated security framework with fuzzy logic assessment for cultural heritages," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1843–1847, 2011.

[20] Y. Weng, H. Zhou, and J. Wan, "Image inpainting technique based on smart terminal: A case study in cps ancient image data," *IEEE Access*, vol. 7, pp. 69837–69847, 2019.

[21] K. Ryabinin, M. Kolesnik, A. Akhtamzyan, and E. Sudarikova, "Cyber-physical museum exhibits based on additive technologies, tangible interfaces and scientific visualization," *Scientific Visualization*, vol. 11, no. 4, 2019.

[22] G. Nota and G. Petraglia, "Heritage buildings management: the role of situational awareness and cyber-physical systems," *Journal of Ambient Intelligence and Humanized Computing*, vol. 15, no. 4, 2024.

[23] L. M. Khodeir and H. E. Soliman, "Sustainable development of heritage areas: Towards cyber-physical systems integration in extant heritage buildings and planning conservation," *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, vol. 5, no. 1, pp. 40–53, 2017.

[24] G. Mylonas, A. Kalogeras, G. Pavlidis, A. Lalos, and A. García-López, "Digital twins for protecting cultural heritage against climate change," *Computer*, vol. 56, no. 9, pp. 100–104, 2023.

[25] Author names ommited for peer review purposes. Title ommited for peer review purposes.

[26] D. Branco, A. Amato, S. Venticinque, and R. Aversa, "Agents based cyber-physical diffused museums over web interoperability standards," *IEEE Access*, vol. 11, pp. 44107–44122, 2023.

[27] N. Patrizi, S. K. LaTouf, E. E. Tsiropoulou, and S. Papavassiliou, "Museum and visitor interaction and feedback orchestration enabled by labor economics," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 5, pp. 2870–2881, 2023.

[28] Y. Naudet, B. A. Yilma, and H. Panetto, "Personalisation in cyber physical and social systems: the case of recommendations in cultural heritage spaces," in *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pp. 75–79, IEEE, 2018.

[29] R. Tang, Y.-N. Chuang, and X. Hu, "The science of detecting llm-generated text," *Commun. ACM*, vol. 67, p. 50–59, Mar. 2024.

[30] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE access*, vol. 12, pp. 26839–26874, 2024.

[31] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, vol. 4, no. 2, p. 100211, 2024.

[32] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *ACM Trans. Knowl. Discov. Data*, vol. 18, Apr. 2024.

[33] Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, L. Yu, Y. Liu, J. Li, B. Xiong, D. Xiong, *et al.*, "Evaluating large language models: A comprehensive survey," *arXiv preprint arXiv:2310.19736*, 2023.

[34] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, and Y. Liu, "Llms-as-judges: A comprehensive survey on llm-based evaluation methods," 2024.

[35] S. Varitimiadis, K. Kotis, A. Skamagis, A. Tzortzakakis, G. Tsekouras, and D. Spiliotopoulos, "Towards implementing an AI chatbot platform for museums," in *International conference on cultural informatics, communication & media studies*, vol. 1, 2020.

[36] J. Jeon, S. Lee, and H. Choe, "Beyond chatgpt: A conceptual framework and systematic review of speech-recognition chatbots for language learning," *Computers & Education*, p. 104898, 2023.

[37] G. Yenduri, M. Ramalingam, G. C. Selvi, Y. Supriya, G. Srivastava, P. K. R. Maddikunta, G. D. Raj, R. H. Jhaveri, B. Prabadevi, W. Wang, *et al.*, "Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," *IEEE Access*, 2024.

[38] W. Xu, M. Liu, O. Sokolsky, I. Lee, and F. Kong, "Llm-enabled cyber-physical systems: survey, research opportunities, and challenges," in *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*, pp. 50–55, IEEE, 2024.

[39] D. Branco, A. Amato, S. Venticinque, and R. Aversa, "Agents based cyber-physical diffused museums over web interoperability standards," *IEEE Access*, vol. 11, pp. 44107–44122, 2023.

[40] P. A. Jansen, "Visually-grounded planning without vision: Language models infer detailed plans from high-level instructions," *arXiv preprint arXiv:2009.14259*, 2020.

[41] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, "Text2motion: From natural language instructions to feasible plans," *Autonomous Robots*, vol. 47, no. 8, pp. 1345–1365, 2023.

[42] W. Xu, M. Liu, S. Drager, M. Anderson, and F. Kong, "Assuring llm-enabled cyber-physical systems," in *2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS)*, pp. 287–288, IEEE, 2024.

[43] H. Jin, T. Zhang, A. Ramamurthy, A. Hamza, and M. Malinoski, "Learning to verify and assure cyber-physical systems," in *AIAA SCITECH 2024 Forum*, p. 1853, 2024.

[44] X. Xue, X. Yu, and F.-Y. Wang, "Chatgpt chats on computational experiments: From interactive intelligence to imaginative intelligence for design of artificial societies and optimization of foundational models," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 6, pp. 1357–1360, 2023.

[45] A. Tamekuri and S. Yamaguchi, "Provide interpretability of document classification by large language models based on word masking," *Journal of Information Processing*, vol. 32, pp. 466–470, 2024.

[46] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, *et al.*, "Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects," *Authorea Preprints*, vol. 1, pp. 1–26, 2023.

[47] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, Jan. 2025.

[48] A. Bavaresco, R. Bernardi, L. Bertolazzi, D. Elliott, R. Fernández, A. Gatt, E. Ghaleb, M. Giulianelli, M. Hanna, A. Koller, A. F. T. Martins, P. Mondorf, V. Neplenbroek, S. Pezzelle, B. Plank, D. Schlangen, A. Suglia, A. K. Surikuchi, E. Takmaz, and A. Testoni, "Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks," 2024.

[49] Y. Dubois, C. X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. S. Liang, and T. B. Hashimoto, "Alpacafarm: A simulation framework for methods that learn from human feedback," *Advances in Neural Information Processing Systems*, vol. 36, pp. 30039–30069, 2023.

[50] X. Xu, K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli, "An llm can fool itself: A prompt-based adversarial attack," 2023.

[51] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin, "End-to-end open-domain question answering with BERTserini," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (W. Am-

mar, A. Louis, and N. Mostafazadeh, eds.), (Minneapolis, Minnesota), pp. 72–77, Association for Computational Linguistics, June 2019.

[52] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (P. Merlo, J. Tiedemann, and R. Tsarfaty, eds.), (Online), pp. 874–880, Association for Computational Linguistics, Apr. 2021.

[53] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, Mar. 2024.

[54] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid, "The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities," *arXiv preprint arXiv:2408.13296*, 2024.

[55] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.

[56] S. Heimersheim, "You can remove gpt2's layernorm by fine-tuning," *arXiv preprint arXiv:2409.13710*, 2024.